Entropy of Mutating Viruses

J. Ostrowski, M. Ozimek, and K.W. Fornalski*

Faculty of Physics, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warszawa, Poland

Received: 10.05.2024 & Accepted: 03.09.2024

Doi: 10.12693/APhysPolA.146.265 *e-mail: krzysztof.fornalski@pw.edu.pl

A virus is a biophysical dissipative system, far from thermodynamic equilibrium, self-organizing, and self-adapting during evolution. This means, in accordance with the second law of thermodynamics developed by Prigogine, that entropy production and entropy flux decrease the internal entropy of the virus. We showed that the Shannon entropy change of three different viruses (SARS-CoV-2, HIV, and influenza) is constant but negative. This confirms Vopson's hypothesis that virus evolution shows a linear negative decrease in information entropy. It also suggests that viruses are evolving systems that are unlikely to reach their stationary (steady) states due to their discrete host-to-host multiplication.

topics: entropy, SARS-CoV-2, HIV, influenza

1. Introduction

A virus is a physical open system, far from thermodynamic equilibrium. This means that local entropy production causes the system entropy to decrease. However, the virus is not a typical biological organism — its evolution towards better adaptation to changing environments is only possible thanks to hosts, in which viruses can replicate and mutate. Therefore, viruses are self-adapting and self-organizing systems in discrete time relationships depending on the host-to-host manner.

Recently, Melvin Vopson and colleagues [1–3] showed that the genetic Shannon information entropy of the SARS-CoV-2 virus decreases with the number of its mutations (which is correlated with time). The main motivation of our work is to verify these calculations and test them for other types of viruses, such as HIV or influenza. This is of crucial importance because other entropy-related methods have been used in the past [4, 5].

The second motivation of our work is to test our hypothesis that the evolution of viruses tends towards minimal entropy production (stationary states). This can be verified by comparing the dynamics of information entropy decrease between a new population virus (SARS-CoV-2), an intermediate one (HIV), and the virus that has been present in the population for a very long time (influenza).

2. Materials and methods

Three types of viral mutation data, namely SARS-CoV-2, HIV, and influenza, were downloaded from the free database of the National Center for Biotechnology Information (NCBI) [6]. The downloaded data are customized to consist of: date, genome length, and full genotype, stored as letters corresponding to nucleotides (elements of DNA). The IUPAC (International Union of Pure and Applied Chemistry) standard for data storage is used. The archive is in the form of a single file with variants separated by blank lines and can later be extracted into files containing information for individual genomes. The database contains 8.8 million entries for SARS-CoV-2, 1.1 million for HIV and 1 million for influenza. Each entry corresponds to a single DNA sequence of the viral genome.

The method of calculating genetic Shannon information entropy was used, as described by Vopson [1–3], in relation to the information in DNA nucleotide sequences. The entropy is thus defined as

$$S = -\sum_{i=1}^{n} p_i \log_m(p_i), \tag{1}$$

where m is the base of logarithm and defines the unit of information entropy, and has been set to 2 to obtain results in bits, and p_i is the probability of each nucleotide in the genome, interpreted as the

number of specific nucleotides in the genotype divided by the total number of the nucleotides in the genome.

To make it clearer, keeping the bit as the unit, consider a genome of 12 exemplary nucleotides: ACTGAACTGACT. Then the entropy value can be calculated as follows

$$S = -\Big[p_{\mathrm{A}} \log_2(p_{\mathrm{A}}) + p_{\mathrm{C}} \log_2(p_{\mathrm{C}}) + p_{\mathrm{T}} \log_2(p_{\mathrm{T}}) + p_{\mathrm{G}} \log_2(p_{\mathrm{G}})\Big] = -\Big[\tfrac{1}{3}\log_2\left(\tfrac{1}{3}\right) + \tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right) + \tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right) + \tfrac{1}{6}\log_2\left(\tfrac{1}{6}\right)\Big] = 1.959. \quad (2)$$

The information entropy values of the whole virus genomes are presented as a function of the number of detected mutations (M), which is assumed to be a function of time (t) (this relationship is difficult to obtain, see Sect. 4). Please note that within this context the entropy presented is related to the nucleotide sequence in the virus DNA. Next, a linear fit, as the most likely one, was used to check if there was a statistically significant decrease in such defined entropy

$$S = A\,M + S_{\mathrm{initial}}, \quad (3)$$

where A is a linear function slope, M is a mutation number, and S_{initial} is the initial value of entropy of the reference virus (where $M = 0$).

Two regression methods were used: classical least squares (CLS) and robust Bayesian regression (RBR) method [7]. The validity of CLS method was verified by the χ^2 test with the number of degrees of freedom. The latter, RBR, is dedicated to the situation when many outliers and a large scatter of data are expected and is described in the Appendix.

3. Results

The Shannon entropy of the nucleotide sequence was calculated and presented as a function of mutation numbers for three viruses: SARS-CoV-2 in Fig. 1, HIV in Fig. 2, and influenza in Fig. 3. A large scatter with many outliers is clearly visible, but a general linear trend seems to be represented. Therefore, a linear relationship from (3) was fitted to all three portions of the data, using two different statistical methods: CLS and RBR. In this paper, we focus on analyzing the linear dependence, which is motivated by results obtained by Vopson. Other non-linear fits are, in our opinion, less likely, and for this reason, will not be analyzed here.

Results of linear fits (slopes A) are presented in Table I, both for the CLS and RBR method. All results show the statistically significant decrease in entropy as a function of mutation numbers — all values of slopes are negative ($A < 0$). The largest value of negative slope was calculated for influenza, while the smallest (but still significant), for HIV virus. Those results simply confirmed the Vopson's hypothesis [1–3].

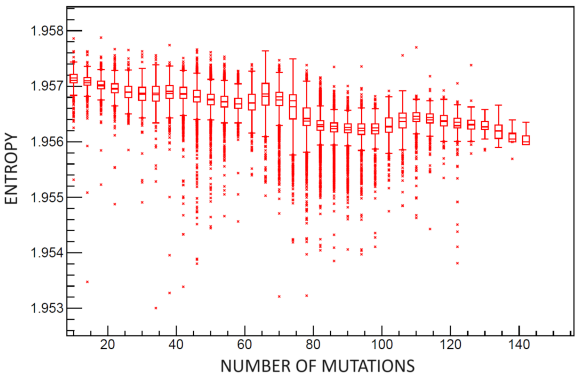

Fig. 1. Information entropy of SARS-CoV-2 virus related to the number of mutations, based on the NCBI database, with boxes closing on the first and third quantile and both the mean and the median marked inside of them. Outer whiskers symbolize furthest away value that is within 1.5 length of a box away from it. Visible points mark outliers that are not within those bounds.

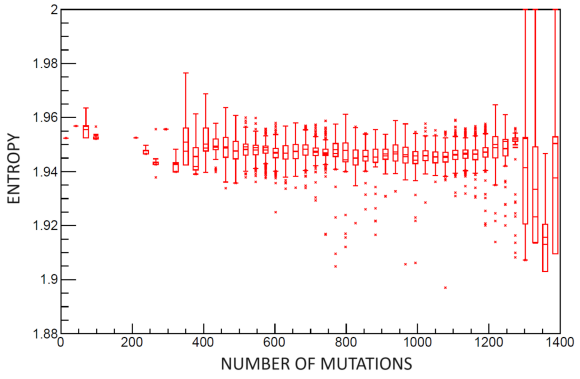

Fig. 2. Information entropy of HIV virus related to the number of mutations, based on the NCBI database, with boxes closing on the first and third quantile and both the mean and the median marked inside of them. Outer whiskers symbolize furthest away value that is within 1.5 length of a box away from it. Visible points mark outliers that are not within those bounds.

Results given by the CLS and RBR methods are consistent — especially because of the large uncertainties in the RBR method. Please note that these methods are based on completely different approaches [7], thus presented results have better likelihood due to the mentioned consistency.

Negative but constant values of A result in a constant entropy reduction, $\mathrm{d}S = A < 0$ (see (3)). Because the entropy change, due to the local form of the second law of thermodynamics [8, 9], is a sum of the entropy production, $\mathrm{d}_i S$ (which is always positive), and the entropy flux, $\mathrm{d}_e S$, the total system's entropy change is therefore given by

$$\mathrm{d}S = \mathrm{d}_e S + \mathrm{d}_i S = A < 0. \quad (4)$$

TABLE I

Results of the best linear fit using classical least squares (CLS) and robust Bayesian regression (RBR) methods. Linear functions (3) were fitted to the genetic information entropy functions of SARS-CoV-2, HIV, and influenza viruses, presented in Fig. 1, Fig. 2, and Fig. 3, respectively. Presented values are equal to the linear slope (A) from (3). All uncertainties represent one standard deviation.

Type of virus	CLS		RBR	Reference figure
SARS-CoV-2	$A = (-0.972 \pm 0.018) \times 10^{-5}$	$\chi^2 = 7.6 \times 10^7$	$A = (-1.146 \pm 0.421) \times 10^{-5}$	Fig. 1
HIV	$A = (-0.223 \pm 0.024) \times 10^{-5}$	$\chi^2 = 1.6 \times 10^9$	$A = (-0.433 \pm 0.338) \times 10^{-5}$	Fig. 2
Influenza	$A = (-1.802 \pm 0.093) \times 10^{-5}$	$\chi^2 = 4.0 \times 10^{10}$	$A = (-1.725 \pm 1.245) \times 10^{-5}$	Fig. 3

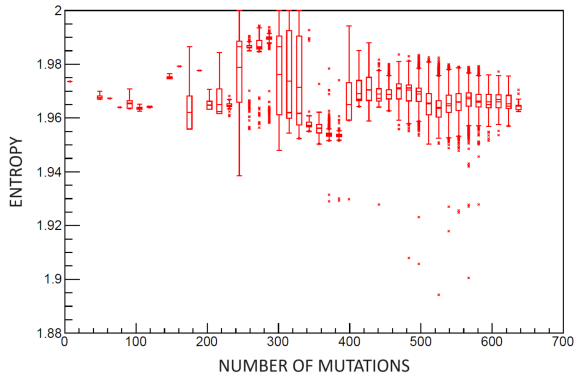

Fig. 3. Information entropy of influenza virus related to the number of mutations, based on the NCBI database, with boxes closing on the first and third quantile and both the mean and the median marked inside of them. Outer whiskers symbolize furthest away value that is within 1.5 length of a box away from it. Visible points mark outliers that are not within those bounds.

So, $\mathrm{d}_e S < -\mathrm{d}_i S < 0$, which is consistent with the Prigogine's and Onsager's theory of the non-equilibrium thermodynamics [8–12]. Thus, the entropy decrease of an evolving system, like a virus, is something natural (assuming that the nucleotide sequence information entropy is representative of the entropy in general).

4. Discussion

The second law of thermodynamics describes the tendency of a thermodynamic system to increase its entropy until it reaches a maximum at equilibrium for all isolated systems [13]. It is important to note that the change in entropy for equilibrium systems is calculated as the difference in entropy between two equilibrium states. This change can occur both in reversible processes (i.e., those that can proceed in the opposite direction, where $\mathrm{d}S = \mathrm{d}Q/T$, and the entropy of the system and the surroundings remains constant) and irreversible processes (where $\mathrm{d}S > \mathrm{d}Q/T$, and the entropy of the system and the surroundings increases) [12]. However, for non-equilibrium state of the system, its entropy can locally decrease, as has been successfully explained by Prigogine and Onsager [9–12], especially for living things. Viruses are a very special case of such a system: it is a self-organized structure, but the evolution processes are narrowed mostly to virus–host interactions.

The modern non-equilibrium thermodynamics exceeded its second law. It allows us to analyze different complex processes like the system multiplication [14], dissipative adaptation in driven self-assembly [15], or the adaptive evolution of complex systems [16]. All this make the second law of thermodynamics a much more useful tool than it was used in the past [8]. This can be applied to analyze the evolution of viruses as well.

To analyze this problem, the simple Shannon information entropy of the sequence of DNA nucleotide symbols was used. We are aware, however, that the Shannon entropy defined in this way is a basic measure of complexity that can be used to estimate the viral diversity and the existence of its bias. But anyway, it was used in other viral studies as well [1–3, 17], so one can assume that this method is simple but accurate. Please note also that the main aim of the presented paper is to show the preliminary results of our studies on the evolution of three selected viruses. These are, in our opinion, very promising and indicate the need to verify other methods as well.

The presented study shows that the viruses' information entropy reduction is constant but negative, i.e., $\mathrm{d}S = A < 0$. The values of the linear slope, A, are always negative and statistically significant, see Table I. This seems to be completely natural due to non-equilibrium thermodynamics. This was originally confirmed by Melvin Vopson and colleagues [1–3], but for a much more limited amount of data for the SARS-CoV-2 virus only. Therefore, the presented analysis is a significant extension of Vopson's findings with deeper thermodynamic explanations.

Usually, far-from-equilibrium evolutionary systems reach their stationary states, in which the entropy change is zero, $\mathrm{d}S = 0$ [10, 11]. This was not observed for three analyzed viruses,

namely SARS-CoV-2 (which is responsible for the COVID-19 disease), HIV (which is responsible for the AIDS disease), and influenza (which is responsible for the flu). It should be noted that historically, the evolution of the influenza virus is much longer than the evolution of the HIV virus. Analogically, the evolution of HIV is longer than that of the SARS-CoV-2 virus, which is the youngest virus among those three cases. However, we cannot see these evolutionary (time) differences when looking at the entropy reductions (A) related to the number of mutations. One can expect that the oldest virus would have A close to zero, while the youngest virus would have a much stronger A. But looking at Table I, one can observe relatively different situation, i.e., that the absolute A value is the highest for influenza and the lowest for HIV. Therefore, the hypothesis that time of evolution makes $A \to 0$ cannot be true in this specific case.

One has to note, however, that only for the SARS-CoV-2 virus we have full genetic information, because this virus (and its mutations) was sequenced from the very beginning. This is not the case for HIV and influenza, which infected humanity before DNA sequencing was discovered. Therefore, this can create a potential bias, which strongly changes results from Table I. The only way is to continue the entropic observation of these viruses and include other types of them.

A more probable explanation is that viruses, as a quasi-living and time-discrete systems, cannot reach their stable stationary state. Viruses are unstable by nature, always mutate and evolve in relatively cyclic way. Those mutations generate fluctuations that are responsible for a globally constant entropy production, according to the fluctuation–dissipation theorem. However, it should be noted that this theorem is suitable for systems close to thermodynamic equilibrium. In our case, viruses can be treated as systems far from equilibrium, and the fluctuation theorem may be discussed in more detail [18]. This hypothesis, however, requires further studies.

5. Conclusions

The presented analysis fully confirmed the results of Vopson and his colleagues [1–3]. The information entropy of nucleotides decreases when the number of mutations increases and this is observed for three independent viruses: SARS-CoV-2, HIV, and influenza. This is consistent with the theories of Prigogine and Onsager, where the self-adapting dissipative systems, far from thermodynamic equilibrium, reduce their entropy values. The reduction of virus information entropy is purely linear, which means that all three viruses did not reach their stationary states. This indicates that, in the case of viruses as well, one can refer to complex systems that are non-stationary and that reduce their entropy at the cost of an increase in the entropy of their environment. The decreasing entropy of such a system leads to the possibility of self-organisation. However, it was impossible to determine whether a relatively young virus, like SARS-CoV-2, has higher entropy decrease than much older ones (HIV, influenza), or not. In our analysis, no stabilization of the entropy decrease at some minimal value was observed; but due to the lack of genetic information about the HIV and influenza first mutations (a long time ago), this comparison is now practically impossible. The presented analysis, however, can help in better prognosis of viruses' mutation directions, which can be potentially helpful for medicine and society.

Appendix: Robust Bayesian regression method

As a result of the conducted analyses, we expect a large number of points significantly deviating from local maxima both on the entropy axis and the number of mutations. This may affect the reliability of the fits obtained by the classical least squares method. To ensure the correctness of global trend results, it was decided to use the Bayesian method, which effectively deals with the problem of outliers and a significant spread of the analyzed data. The algorithm's operation scheme is described below [7]:

- Acquiring experimental data;
- Selecting the fitting model function $T_i(x_i)$ (e.g., polynomial), with initially established parameters λ and variable x (for linear fitting $T_i(x_i) = \lambda_1 + \lambda_2 x_i$);
- Proposing initial parameter values;
- Creating equations for minimizing function variability based on point statistical weights g_i, taking into account the discrepancy between experimental and theoretical data, $R_i = T_i - y_i$, where $T_i(x_i)$ represents the proposed fitting model function;
- Obtaining new values λ' from these equations;
- Comparing the discrepancies of λ and λ' values with the accuracy-defining parameter ϵ, suggesting that discrepancies of one order of magnitude are smaller than a significant digits of λ;
- Obtaining the result or creating new equations (i.e., entering next iteration).

The statistical weight function of a point is defined as

$$g_i = \frac{1}{R_i^2}\left(2 - \frac{\dfrac{R_i^2}{\sigma_{0i}^2}}{\exp\left(\dfrac{R_i^2}{2\sigma_{0i}^2}\right) - 1}\right), \tag{5}$$

where σ_{0i} denotes the original uncertainty of the i-th point. More information about this method can be found in literature [7].

References

[1] M.M. Vopson, S.C. Robson, *Physica A* **584**, 126383 (2021).

[2] M.M. Vopson, *Appl. Sci.* **12**, 6912 (2022).

[3] M.M. Vopson, S. Lepadatu, *AIP Adv.* **12**, 075310 (2022).

[4] K. Pan, M.W. Deem, *J. R. Soc. Interface* **8**, 1644 (2011).

[5] Y. Zhang, K.M. Eskridge, S. Zhang, G. Lu, *BMC Bioinformatics* **23**, 333 (2022).

[6] NCBI (National Center for Biotechnology Information) Database, 2024.

[7] K.W. Fornalski, *Int. J. Soc. Syst. Sci.* **7**, 314 (2015).

[8] I. Prigogine, G. Nicolis, A. Babloyantz, *Phys. Today* **25**, 23 (1972).

[9] K. Michaelian, *Foundations* **2**, 308 (2022).

[10] L. Onsager, *Phys. Rev.* **37**, 405 (1931).

[11] L. Onsager, *Phys. Rev.* **38**, 2265 (1931).

[12] I. Prigogine, *Introduction to Thermodynamics of Irreversible Processes*, 3rd ed., John Wiley & Sons, New York 1967.

[13] K. Huang, *Statistical Mechanics*, 2nd ed., Wiley, New York 1987.

[14] J.L. England, *J. Chem. Phys.* **139**, 121923 (2013).

[15] J.L. England, *Nat. Nanotechnol.* **10**, 919 (2015).

[16] N. Perunov, R.A. Marsland, J.L. England, *Phys. Rev. X* **6**, 021036 (2016).

[17] A. Gall, S. Kaye, S. Hué et al., *Retrovirology* **10**, 8 (2013).

[18] R. Marsland, J. England, *Phys. Rev. E* **92**, 052120 (2015).